WHAT IS CLAIMED IS:

1. A method for selecting epigenetic features, comprising the steps of:

a) collecting and storing biological samples containing genomic DNA;

b) collecting and storing available phenotypic information about the biological samples so as to define a phenotypic data set;

c) defining at least one phenotypic parameter of interest;

d) dividing the biological samples into at least two disjunct phenotypic classes of interest using the defined phenotypic parameters of interest;

e) defining an initial set of epigenetic features of interest;

f) analysing the defined epigenetic features of interest of the biological samples so as to generate an epigenetic feature data set;

g) selecting relevant epigenetic features of interest and/or combinations of epigenetic features of interest of the defined epigenetic features of interest, the relevant epigenetic features of interest and/or combinations of epigenetic features of interest being relevant for epigenetically-based prediction of the at least two phenotypic classes of interest; and

h) defining a new set of epigenetic features of interest based on the relevant epigenetic features of interest and/or combinations of epigenetic features of interest generated in step g).

2. The method as recited in claim 1 further comprising repeating steps f) and g) based on the new set of epigenetic features of interest defined in step h).

3. The method as recited in claim 1 wherein the biological samples include at least one of cells, cellular components which contain DNA, sources of DNA, tissue embedded in paraffin and histologic object slides.

4. The method as recited in claim 3 wherein the sources of DNA include at least one of cell lines, biopsies, blood, sputum, stool, urine and cerebral-spinal fluid.

5. The method as recited in claim 3 wherein the tissue embedded in paraffin includes at least one of tissue from eyes, intestine, kidney, brain, heart, prostate, lung, breast and liver.

6. The method as recited in claim 1 wherein at least one of the phenotypic information and the phenotypic parameter of interest are selected from the group consisting of kind of tissue, drug resistance, toxicology, organ type, age, life style, disease history, signaling chains, protein synthesis, behavior, drug abuse, patient history, cellular parameters, treatment history and gene expression and combinations thereof.

7. The method as recited in claim 1 wherein the epigenetic features of interest include cytosine methylation sites in DNA.

8. The method as recited in claim 1 wherein the initial set of epigenetic features of interest is defined using preliminary knowledge data about their correlation with phenotypic parameters.

9. The method as recited in claim 1 wherein the relevant epigenetic feature or a combination of epigenetic features is relevant for epigenetically-based prediction of said phenotypic classes of interest when at least one of an accuracy and a significance of the epigenetically-based prediction of the phenotypic classes of interest is likely to decrease by exclusion of the corresponding epigenetic feature data of the epigenetic feature data set.

10. The method as recited in claim 1 wherein step d) is performed so as to divide the biological samples in two disjunct phenotypic classes of interest.

11. The method as recited in claim 10 further comprising performing the epigenetically-based prediction of the at least two phenotypic classes of interest using a machine learning classifier.

12. The method as recited in claim 1 further comprising:

selecting pairs of classes or pairs of unions of classes from the disjunct phenotypic classes of interest; and

performing epigenetically-based prediction of each pair of classes or pair of unions of classes using a machine learning classifier.

13. The method as recited in claim 11 wherein the selecting of step g) includes:

defining a candidate set of and/or combinations of epigenetic features of interest of the defined epigenetic features of interest;

defining a feature selection criterion;

ranking the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest according to the feature selection criterion; and

selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

14. The method as recited in claim 13 wherein the candidate set of epigenetic features of interest is the set of all subsets of the defined epigenetic features of interest.

15. The method as recited in claim 13 wherein the candidate set of epigenetic features of interest is a set of all subsets of a given cardinality of the defined epigenetic features of interest.

16. The method as recited in claim 13 wherein the candidate set of epigenetic features of interest is a set of all subsets of cardinality 1 of the defined epigenetic features of interest.

17. The method as recited in claim 13 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to principal component analysis, principal components of the principal component analysis defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

18. The method as recited in claim 13 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to multidimensional scaling, calculated coordinate vectors of the multidimensional scaling defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

19. The method as recited in claim 13 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by sub-

jecting the epigenetic feature data set to isometric feature mapping, calculated coordinate vectors of the isometric feature mapping defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

20. The method as recited in claim 13 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to cluster analysis and then combining epigenetic features of interest belonging to a same cluster so to define said candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

21. The method as recited in claim 20 wherein the cluster analysis includes hierarchical clustering.

22. The method as recited in claim 20 wherein the cluster analysis includes k-means clustering.

23. The method as recited in claim 13 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed using predetermined biological information.

24. The method as recited in claim 23 wherein the biological information includes at least one biological factor selected from the group consisting of: correlated methylation status, proximity of epigenetic features to each other on a genome, epigenetic features located on a same gene, epigenetic features that are a exon/intron/promoter of a same gene, epigenetic features located on genes that are co-regulated, epigenetic features located on genes that have similar biological functionality, and epigenetic features located on genes that are part of the same biological pathway.

25. The method as recited in claim 13 wherein the feature selection criterion includes a training error of the machine learning classifier trained on respective epigenetic feature data of the epigenetic feature data set corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

26. The method as recited in claim 13 wherein the feature selection criterion includes a risk of the machine learning classifier trained on epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

27. The method as recited in claim 13 wherein the feature selection criterion includes a bound on a risk of the machine learning classifier trained on epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

28. The method as recited in claim 13 wherein the feature selection criterion includes a statistical test for computing a significance of difference of the phenotypic classes of interest given epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

29. The method as recited in claim 28 wherein the statistical test includes a t-test.

30. The method as recited in claim 28 wherein the statistical test includes a rank test.

31. The method as recited in claim 30 wherein the rank test includes a Wilcoxon rank test.

32. The method as recited in claim 28 wherein the statistical test includes a multivariate test.

33. The method as recited in claim 32 wherein the multivariate test includes a T^2-test.

34. The method as recited in claim 32 wherein the multivariate test includes a likelihood ratio test for logistic regression models.

35. The method as recited in claim 13 wherein the feature selection criterion includes a Fisher criterion for the phenotypic classes of interest given the epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

36. The method as recited in claim 13 wherein the feature selection criterion includes weights of a linear discriminant for the phenotypic classes of interest given epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

37. The method as recited in claim 36 wherein the linear discriminant is a Fisher discriminant.

38. The method as recited in claim 36 wherein the linear discriminant is a discriminant of a support vector machine classifier for the phenotypic classes of interest trained on epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

39. The method as recited in claim 13 wherein the defining the epigenetic feature selection criterion includes subjecting epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest to principal component analysis and calculating weights of a first principal component.

40. The method as recited in claim 13 wherein the epigenetic feature selection criterion includes an average pairwise correlation between all single features in a given subset of epigenetic features on a given set of samples.

41. The method as recited in claim 13 wherein the epigenetic feature selection criterion includes mutual information between the phenotypic classes of interest and a classification achieved by an optimally selected threshold on s given epigenetic feature of interest.

42. The method as recited in claim 13 wherein the epigenetic feature selection criterion includes a number of correct classifications achieved by an optimally selected threshold on a given epigenetic feature of interest.

43. The method as recited in claim 13 wherein the epigenetic feature selection criterion includes eigenvalues of the principal components.

44. The method as recited in claim 13 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting a defined number of highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

45. The method as recited in claim 13 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting all except a defined number of lowest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

46. The method as recited in claim 13 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold.

47. The method as recited in claim 13 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score lower than a defined threshold.

48. The method as recited in claim 2 wherein the repeating steps f) and g) is performed until a defined number of the epigenetic features of interest and/or combinations of epigenetic features of interest are selected.

49. The method as recited in claim 2 wherein the repeating steps f) and g) is performed until all epigenetic features of interest and/or combinations of epigenetic features of interest of the epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected.

50. The method as recited in claim 2 further comprising determining an optimal number of epigenetic features of interest and/or combinations of epigenetic features of interest using crossvalidation of a machine learning classifier on test subsets of epigenetic feature data.

51. The method as recited in claim 13 further comprising determining an optimal feature selection criterion score threshold by crossvalidation of the classifier on test subsets of epigenetic feature data.

52. The method as recited in claim 1 further comprising training a machine learning classifier using a feature data set corresponding to the defined new set of epigenetic features of interest.

53. A computer readable medium having stored thereon computer executable process steps operative to perform a method for selecting epigenetic features, the method comprising the steps of:

a) receiving an epigenetic feature data set for a plurality of epigenetic features of interest, the epigenetic feature data set being grouped in disjunct classes of interest;

b) selecting relevant epigenetic features of interest and/or combinations of epigenetic features of interest of the plurality of epigenetic features of interest, the relevant epigenetic features of interest and/or combinations of epigenetic features of interest being relevant for machine learning class prediction based on corresponding epigenetic feature data of the epigenetic feature data set; and

c) defining a new set of epigenetic features of interest based on the relevant epigenetic features of interest and/or combinations of epigenetic features of interest generated in step b).

54. The computer readable medium as recited in claim 53 wherein the method further comprises repeating step b) based on the new set of epigenetic features of interest defined in step c).

55. The computer readable medium as recited in claim 53 wherein the relevant epigenetic features of interest and/or combinations of epigenetic features of interest are relevant for machine learning class prediction when at least one of an accuracy and a significance of the machine learning class prediction is likely to decrease by exclusion of the corresponding epigenetic feature data.

56. The computer readable medium as recited in claim 53 wherein the method further comprises grouping the epigenetic feature data set in disjunct pairs of classes and/or pairs of unions of classes of interest before performing steps b) and c).

57. The computer readable medium as recited in claim 53 wherein the selecting of step b) includes:

defining a candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest of the plurality of epigenetic features of interest;

defining a feature selection criterion;

ranking the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest according to the feature selection criterion; and

selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

58. The computer readable medium as recited in claim 57 wherein the candidate set of epigenetic features of interest is the set of all subsets of the defined epigenetic features of interest.

59. The computer readable medium as recited in claim 57 wherein the candidate set of epigenetic features of interest is a set of all subsets of a given cardinality of the defined epigenetic features of interest.

60. The computer readable medium as recited in claim 57 wherein the candidate set of epigenetic features of interest is a set of all subsets of cardinality 1 of the defined epigenetic features of interest.

61. The computer readable medium as recited in claim 57 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to principal component analysis, principal components of the principal component analysis defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

62. The computer readable medium as recited in claim 57 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to multidimensional scaling, calculated coordinate vectors of the multidimensional scaling defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

63. The computer readable medium as recited in claim 57 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to isometric feature mapping, calculated coordinate vectors of the isometric feature mapping defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

64. The computer readable medium as recited in claim 57 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed by subjecting the epigenetic feature data set to cluster analysis and then combining epigenetic features of interest belonging to a same cluster so to define said candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

65. The computer readable medium as recited in claim 64 wherein the cluster analysis includes hierarchical clustering.

66. The computer readable medium as recited in claim 64 wherein the cluster analysis includes k-means clustering.

67. The computer readable medium as recited in claim 57 wherein the defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is performed using predetermined biological information.

68. The computer readable medium as recited in claim 67 wherein the biological information includes at least one biological factor selected from the group consisting of: correlated methylation status, proximity of epigenetic features to each other on a genome, epigenetic

features located on a same gene, epigenetic features that are a exon/intron/promoter of a same gene, epigenetic features located on genes that are co-regulated, epigenetic features located on genes that have similar biological functionality, and epigenetic features located on genes that are part of the same biological pathway.

69. The computer readable medium as recited in claim 57 wherein the feature selection criterion includes a training error of the machine learning classifier trained on respective epigenetic feature data of the epigenetic feature data set corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

70. The computer readable medium as recited in claim 57 wherein the feature selection criterion includes a risk of the machine learning classifier trained on epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

71. The computer readable medium as recited in claim 57 wherein the feature selection criterion includes a bound on a risk of the machine learning classifier trained on epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

72. The computer readable medium as recited in claim 57 wherein the feature selection criterion includes a statistical test for computing a significance of difference of the phenotypic classes of interest given epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

73. The computer readable medium as recited in claim 72 wherein the statistical test includes a t-test.

74. The computer readable medium as recited in claim 72 wherein the statistical test includes a rank test.

75. The computer readable medium as recited in claim 74 wherein the rank test includes a Wilcoxon rank test.

76. The computer readable medium as recited in claim 72 wherein the statistical test includes a multivariate test.

77. The computer readable medium as recited in claim 76 wherein the multivariate test includes a T^2-test.

78. The computer readable medium as recited in claim 76 wherein the multivariate test includes a likelihood ratio test for logistic regression models.

79. The computer readable medium as recited in claim 57 wherein the feature selection criterion includes a Fisher criterion for the phenotypic classes of interest given the epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

80. The computer readable medium as recited in claim 57 wherein the feature selection criterion includes weights of a linear discriminant for the phenotypic classes of interest given epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

81. The computer readable medium as recited in claim 80 wherein the linear discriminant is a Fisher discriminant.

82. The computer readable medium as recited in claim 80 wherein the linear discriminant is a discriminant of a support vector machine classifier for the phenotypic classes of interest trained on epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

83. The computer readable medium as recited in claim 57 wherein the defining the epigenetic feature selection criterion includes subjecting respective epigenetic feature data of the epigenetic feature data set corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest to principal component analysis and calculating weights of a first principal component.

84. The computer readable medium as recited in claim 57 wherein the epigenetic feature selection criterion includes an average degree of methylation on a set of samples with given phenotypical properties.

85. The computer readable medium as recited in claim 57 wherein the epigenetic feature selection criterion includes an average pairwise correlation between all single features in a given subset of epigenetic features on a given set of samples.

86. The computer readable medium as recited in claim 57 wherein the epigenetic feature selection criterion includes mutual information between the phenotypic classes of interest and a classification achieved by an optimally selected threshold on s given epigenetic feature of interest.

87. The computer readable medium as recited in claim 57 wherein the epigenetic feature selection criterion includes a number of correct classifications achieved by an optimally selected threshold on a given epigenetic feature of interest.

88. The computer readable medium as recited in claim 57 wherein the epigenetic feature selection criterion includes eigenvalues of the principal components.

89. The computer readable medium as recited in claim 57 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting a defined number of highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

90. The computer readable medium as recited in claim 57 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting all except a defined number of lowest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

91. The computer readable medium as recited in claim 57 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of inter-

est is performed by selecting epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold.

92. The computer readable medium as recited in claim 57 wherein the selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is performed by selecting epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score lower than a defined threshold.

93. The computer readable medium as recited in claim 54 wherein the repeating step b) is performed until a defined number of the epigenetic features of interest and/or combinations of epigenetic features of interest are selected.

94. The computer readable medium as recited in claim 54 wherein the repeating step b) is performed until all epigenetic features of interest and/or combinations of epigenetic features of interest of the epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected.

95. The computer readable medium as recited in claim 54 wherein the method further comprises determining an optimal number of epigenetic features of interest and/or combinations of epigenetic features of interest of the epigenetic features of interest and/or combinations of epigenetic features of interest using crossvalidation of a machine learning classifier on test subsets of epigenetic feature data.

96. The computer readable medium as recited in claim 54 wherein the method further comprises determining an optimal feature selection criterion score threshold by crossvalidation of a machine learning classifier on test subsets of epigenetic feature data.

97. The computer readable medium as recited in claim 54 wherein the method further comprises training a machine learning classifier using a feature data set corresponding to the defined new set of epigenetic features of interest.